Research Paper | Computer Science | E-ISSN No : 2454-9916 | Volume : 2 | Issue : 6 | June 2016

A SYSTEM TO ANALYSIS REAL TIME BIGDATA USING TOPDOWN SPECIALIZATION

P. Shiva [1] | Dr. P. Karthiga [2] | Dr. X. J. J. Anitha [2]

[1] MCA., Nadar Mahjana Sangam S.Vellaichamy Nadar College.

[2] M.Phil., Ph.D., Nadar Mahjana Sangam S.Vellaichamy Nadar College.

ABSTRACT

The large number of cloud services such as like Application,Platform,Infrastructure (IaaS,PaaS,SaaS)requires the users want to share the private data like for data analysis or data mining, data anonymization, bringing privacy concerns privacy may be data sets via generalization to satisfy certain security and privacy requirements such as k-anonymity and store them categorized of preserving techniques. At present nowadays the data cloud applications are increasing their large-scale data within the Big data trend huge amount of data sets and preserving sensitive, large scale data is very difficult data sets due to their map reduce is design by two phase of this technique to archieve scalable two phase Top Down Specifications (TDS) is scalability and efficient (TDS) is significance improved over existing approaches. TheMap reduce approach is a framework and this widely adopted for parallel data processing to address the scalability problem of the top-down specialization(TDS) approach for large scale data Anonymization. TDS approach is widely used for data Anonymization that provides a good arbitrate between data utility and data consistency. Most of the TDS algorithm is centralized, that are insufficient to handle large-scale data sets. we introduce a highly scalable two phase TDS approach for data Anonymization by using the map reduce frame.

KEY WORDS: Cloud, data anonymization, data partition, privacy preservation,Top Down Specifications (TDS).

I.INTRODUCTION

Recently, a great deal of interest in the field of Big Data and its analysis has risen[1]–[3], The 3Vs have been expanded to other complementary characteristics of big data:

- Volume: big data doesn't sample; it just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
- Machine Learning: big data often doesn't ask why and simply detects patterns.
- Digital footprint: big data is often a cost-free byproduct of digital interaction.

Mainly driven from wide number of research challenges powerfully related to massive applications, such as modeling, processing, querying, mining, and distributing large-scale repositories. The term Big Data classifies specific kinds of data sets comprising formless data, which dwell in data layer of technical computing applications and the Web.

II. Application Scenarios

The data stored in the underlying layer of all these technical computing application scenarios have some precise individualities in common, such as 1) large scale data, which refers to the size and the data warehouse; 2) scalability issues, which refer to the application's likely to be running on large scale (e.g., Big Data); 3) Endure extraction transformation loading (ETL) method from low, raw data to well thought-out data up to certain extent; and 4) development of uncomplicated interpretable analytical over Big Data warehouses[4]-[6] with a view to deliver an intelligent and momentous knowledge for them. The highly scalable two-phase TDS approach for data anonymization based on Map Reduce on cloud. To make full use of the parallel capability of Map Reduce on cloud, specializations required in an anonymization process are split into two phases. In the first one, original datasets are partitioned into a group of smaller datasets, and these datasets are anonymized in parallel, producing intermediate results. In the second one, the intermediate results are integrated into one, and further anonymized to achieve consistent k-anonymous data sets. It leveragesMap Reduce to accomplish the concrete computation in both phases. A group of Map Reduce [7] jobs are deliberately designed and coordinated to perform specializations on data sets collaboratively. It evaluate the approach by conducting experiments on real-world data sets. Experimental results demonstrate that with the approach, the scalability and efficiency of TDS can be improved. It evaluates the approach by conducting experiments on real-world data sets. Experimental results demonstrate that with the approach, the scalability and efficiency of TDS can be improved. Significantly over existing approaches. The major contributions of the research are threefold. Firstly, it creatively apply MapReduce[8] on cloud to TDS for data anonymization and deliberately design a group of innovative Map Reduce jobs to concretely accomplish the specializations in a highly scalable fashion. Secondly, it propose a two-phase TDS approach to gain high scalability via allowing specializations to be conducted on multiple data partitions in parallel during the first phase.

Big Data are usually generated by online transaction, video/audio, email, number of clicks, logs, posts, social network data, scientific data, remote access sensory data, mobile phones, and their applications. These data are accumulated in databases that grow extraordinarily and become complicated to confine, form, store, manage, share, process, analyze, and visualize via typical database software tools.

Advancement in Big Data sensing and computer technology revolutionizes the way remote data collected, processed, analyzed, and managed [9]–[12]. Particularly, most recently designed sensors used in the earth and planetary observatory system are generating continuous stream of data. Moreover, majority of work have been done in the various fields of remote sensory satellite image data, such as change detection, gradient-based edge detection [14], region similarity based edge detection [15], and intensity gradient technique for efficient intra prediction. In this paper, we referred the high speed continuous stream of data or high volume offline data to Big Data, which is leading us to a new world of challenges Such consequences of transformation of remotely sensed data to the scientific understanding are a critical task .. Hence the rate at which volume of the remote access data is increasing, a number of individual users as well as organizations are now demanding an efficient mechanism to collect, process, and analyze, and store these data and its resources.

Big Data analysis is somehow a challenging task than locating, identifying, understanding, and citing data. Having a large-scale data, all of this has to happen in a mechanized manner since it requires diverse data structure as well as semantics to be articulated in forms of computer-readable format. However, by analyzing simple data having one data set, a mechanism is required of how to design a database. There might be alternative ways to store all of the same information. In such conditions, the mentioned design might have an advantage over others for certain process and possible drawbacks for some other purposes. In order to address these needs, various analytical platforms have been provided by relational databases vendors. These platforms come in various shapes from software only to analytical services that run in third-party hosted environment.

III. Network Sensors

In remote access networks, where the data source such as sensors can produce an overwhelming amount of raw data. We refer it to the first step, i.e., data acquisition, in which much of the data are of no interest that can be filtered or compressed by orders of magnitude. With a view to using such filters, they do not discard useful information. For instance, in consideration of new reports, is it adequate to keep that information that is mentioned with the company name Alternatively, is it necessary that we may need the entire report, or simply a small piece around the mentioned name. The second challenge is by default generation of accurate metadata that describe the composition of data and the way it was collected and analyzed. Such kind of metadata is hard to analyze since we may need to know the source for each data in remote access.

Normally, the data collected from remote areas are not in a format ready for analysis. Therefore, the second step refers us to data extraction, which drags out the useful information from the underlying sources and delivers it in a structured formation suitable for analysis. For instance, the data set is reduced to single-class label to facilitate analysis, even though the first thing that we used to think about Big Data as always describing the fact. However, this is far away from reality, sometimes we have to deal with erroneous data too, or some of the data might be imprecise. To address the aforementioned needs, this paper presents a remote sensing Big Data analytical architecture, which is used to analyze real time, as well as offline data. At first, the data are remotely preprocessed, which is then readable by the machines. Afterward, this useful information is transmitted to the Earth Base Station for further data processing.

IV. Station Performance

Earth Base Station performs two types of processing, such as processing of real-time and offline data. In case of the offline data, the data are transmitted to offline data-storage device. The incorporation of offline data-storage device helps in later usage of the data, whereas the real-time data is directly transmitted to the filtration and load balancer server, where filtration algorithm is employed, which extracts the useful information from the Big Data. On the other hand, the load balancer balances the processing power by equal distribution of the real-time data to the servers. The filtration and load-balancing server not only filters and balances the load, but it is also used to enhance the system efficiency.

Furthermore, the filtered data are then processed by the parallel servers and are sent to data aggregation unit (if required, they can store the processed data in the result storage device) for comparison purposes by the decision and analyzing server. The proposed architecture welcomes remote access sensory data as well as direct access network data (e.g., GPRS, 3G, xDSL, or WAN). The proposed architecture and the algorithms are implemented in Hadoop using Map Reduce programming by applying remote sensing earth observatory data.

V.Remote Sensing Big Data Analytics

Two-PhaseTop-Down Specialization (TPTDS) approach to conduct the computation required in TDS in a highly scalable and efficient fashion. The two phases of the approach are based on the two levels of parallelization provisioned by Map Reduce on cloud. Basically, Map Reduce on cloud has two levels of parallelization, i.e., job level and task level. Job level parallelization means that multiple Map Reduce jobs can be executed simultaneously to make full use of cloud infrastructure resources. Combined with cloud, Map Reduce becomes more powerful and elastic as cloud can offer infrastructure resources on demand, e.g., Amazon Elastic Map Reduce service. Task level parallelization refers to that multiple mapper/reducer tasks in a Map Reduce job are executed simultaneously over data splits. To achieve high scalability, parallelizing multiple jobs on data partitions in the first phase, but the resultant anonymization levels are not identical. To obtain finally consistent anonymous data sets, the second phase is necessary to integrate the intermediate results and further anonymize entire data sets. Details are formulated as follows. All intermediate anonymization levels are merged into one in the second phase.

The merging of anonymization levels is completed by merging cuts. Specifically, let in and in be two cuts of an attribute. There exist domain values and that satisfy one of the three conditions is identical to is more general than is more specific than. To ensure that the merged intermediate anonymization level never violates privacy requirements, the more general one is selected as the merged one, e.g., will be selected if is more general than or identical to . For the case of multiple anonymization levels, it can merge them in the same way iteratively. The following lemma ensures that still complies privacy requirements.

The increase in the data rates generated on the digital universe is escalating exponentially. With a view in employing current tools and technologies to analyze and store, a massive volume of data are not up to the mark, since they are unable to extract required sample data sets. Therefore, we must design an architectural platform for analyzing both remote access real time and offline data. When a business enterprise can pull-out all the useful information obtainable in the Big Data rather than a sample of its data set, in that case, it has an influential benefit over the market competitors. Big Data analytics helps us to gain insight and make better decisions. Therefore, with the intentions of using Big Data, modifications in paradigms are at utmost. To support our motivations, we have described some areas where Big Data can play an important role. Understanding environment requires massive amount of data collected from various sources, such as remote access satellite observing earth characteristics [measurement data set (MDS) of satellite data such as images], sensors monitoring air and water quality, metrological circumstances, and proportion of CO2 and other gases in air, and so on. Through relating all the information drifting such as CO2 emanation, increase or decrease on greenhouse effects and temperature, can be found out.

VI. PROPOSED WORK

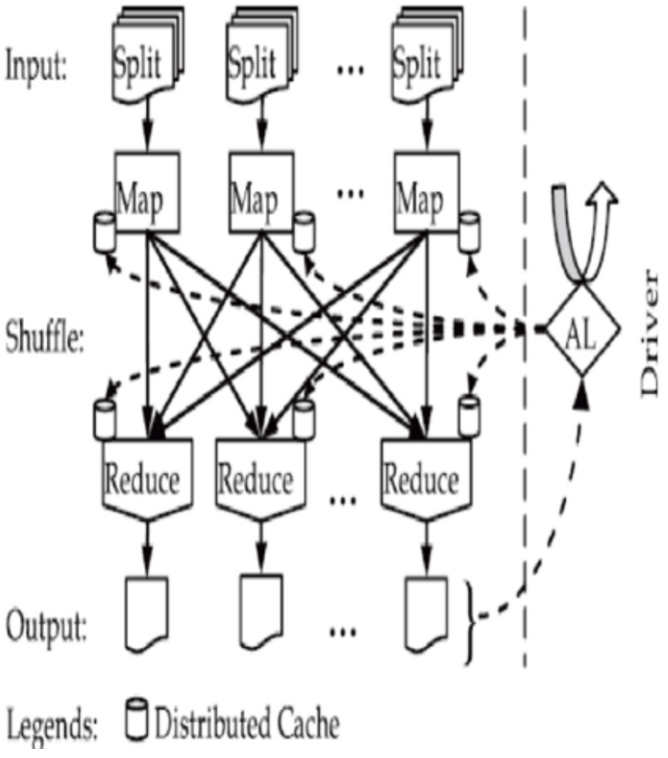
Distributed systems allow for greater overall service performance than systems whose function is centralized in a single location. Instead of storing all kind of data into the same location, we can split the based on their categories. By spreading the computational load across different nodes, each location is under less stress. This allows each node to perform more efficiently, which increases the performance of the overall service. When computation is centered on a single machine, the health of that machine is the health of the entire service, if it goes down, so does the entire service. However, distributed systems can continue to function if one node ceases to function. Because distributed systems work across a variety of different machines, they are inherently scalable. That is, the distributed system can adjust how many system resources it is making use of in light of what kind of demand the system is under.When services run on a single server, there is no worry about data synchronization: all the data is simply present on that machine.In this paper, we propose a scalable two-phase top-down specialization (TDS) approach to secure large-scale data sets using the MapReduce framework on cloud.In both phases of our approach, we deliberately design a group of innovative MapReduce jobs to concretely accomplish the specialization computation in a highly scalable way.This approach gets input data's and split into the small data sets. Then we apply the ANONYMIZATION on small data sets to get intermediate result. Then small data sets are merging and again apply the ANONYMIZATION. We analyze the each and every data set sensitive field and give priority for this sensitive field. Then we apply ANONYMIZATION on this sensitive field only depending upon the scheduling.

VII. ARCHITECTURE

The term Big Data covers diverse technologies same as cloud computing. The input of Big Data comes from social networks, Web servers, satellite imagery, sensory data, banking transactions, etc. Regardless of very recent emergence of Big Data architecture in scientific applications, numerous efforts toward Big Data analytics architecture can already be found in the literature. Among numerous others, we propose remote sensing Big Data architecture to analyze the Big Data in an efficient manner. It delineates n number of satellites that obtain the earth observatory Big Data images with sensors or conventional cameras through which sceneries are recorded using radiations. Special techniques are applied to process and interpret remote sensing imagery for the purpose of producing conventional maps, thematic maps, resource surveys, etc. We have divided remote sensing Big Data architecture into three parts, i.e., 1) remote sensing data acquisition unit (RSDU); 2) data processing unit (DPU); and 3) data analysis and decision unit (DADU). The functionalities and working of the said parts are described as below.

To evaluate the effectiveness and efficiency of our two phase approach, we compare it with the centralized TDS approach proposed in, denoted as CentTDS. CentTDS is the state-of-the-art approach for TDS anonymization. Scalability and data utility are considered for the effectiveness. For scalability, we check whether both approaches can still work and scale over large-scale data sets. Data utility is measured by the metric I Loss, a general purpose data metric proposed. Literally, I Loss means information loss caused by data anonymization. Basically, higher I Loss indicates less data utility. How to calculate I Loss can be found in Appendix A.2, which is available in the online supplemental material. The I Loss of CentTDS and TPTDS are denoted as ILCent and ILTP, respectively. The execution time of CentTDS and TPTDS are denoted as TCent and TTP, respectively.

Map Reduce Architecture



VIII. ALGORITHM USED

Algorithm I. Filtration and Load Balancing Algorithm(FLBA)

Input: Satellite process data set/product

Output: filtered Image data in fixed size block and send each block to processing server

Steps:

1. Filter Image related data i.e. Processed data in MDS. All other unnecessary data will be discarded.
2. Divide the image into fixed size block i.e. BS = 100×100 MDS process_data values, row by row fashion or column by column. Each block will be denoted byBi where1≤I ≤BS
3. Make two samples of blocks so that only half of the part is processed. i.e.,PSB={B1,B3,B5,...,BN−1}andUPSB={B2,B4,B6,B8,...,B}
4. Transmit UPSB directly to aggregation server without processing.
5. Assign and transmit each distinct block(s) Bi of PSB to various processing servers in DPU.

Algorithm II.Processing and Calculation Algorithm (PCA)

Input: BlockBi

Output: statistical parameters results and transmit them toaggregation server.

Steps:

1. For each BlockBi, Calculate
   a. Xbi
   b. S.DBi
   c. Abs_Diff
   d. Ngmaxval
2. Transmit the results against block id and product id to the aggregation server in DADU

Algorithm III.Aggregation and Compilation Algorithm(ACA)

Input: BlockBi results

Output: compiling, storing and sending PSB results and UPSB blocks information to decision-making server.

Steps:

1. Collect EveryBi's result of PSB
2. Compile them and transmit them to Decision-making server.
3. Store PSB blocks with results and UPSB blocks without result into RBMS in result storage.

Algorithm IV.Decision-making algorithm (DMA)

Input: PSB results and UPSB information

Output: each block Bi with decision, land block or sea. Finally, the whole image is divided into sea and land area

Rules:

Following rules are made on the basis of land area analysis discussed in Section III for detecting land block

1. XBi≤∂X
2. S.DBi≥∂S.D
3. Abs_Diff≥∂Abs_diff
4. NGmaxval≤∂NGmaxval

Steps:

1. For Each (Bi of PBS)

{

If (Rule1==true and Rule2==true)

Status_Bi =Land

Else if (Rule1==false and Rule2==false)

Status_Bi =Sea

Else

{

If (Rule3==false and Rule4==false))

Status_Bi =Sea

Else

Status_Bi =Land

}

}

2. For Each (Bi of UPBS)

{

If (Status_Bi−1== Landand status_Bi+1 =Land)

Status_Bi =Land

Else If (Status_Bi−1== Seaand status_Bi+1 =Sea)

Status_Bi =Sea

Else

Status_Bi =! (Status_Bi−1⊕status_Bi+1 ⊕status_Bi+3)

}

IX. CONCLUSION

In this paper, the proposed architecture for a system to analysis real time big data using Top Down Specialization. The proposed architecture efficiently processed and analyzed real-time and offline remote sensing Big Data for decision-making. Instead of storing all kind of data into the same location, we can split the based on their categories. So it will provide quick response time for providing data, no computation overhead and security of data. The proposed architecture is composed of three major units, such as 1) RSDU; 2) DPU; and 3) DADU. These units implement algorithms for each level of the architecture depending on the required analysis. The architecture of real-time Big is generic (application independent) that is used for any type of remote sensing Big Data analysis. Furthermore, the capabilities of filtering, dividing, and parallel processing of only useful information are performed by discarding all other extra data. These processes make a better choice for real-time remote sensing Big Data analysis. The algorithms proposed in this paper for each unit and subunits are used to analyze remote sensing data sets, which helps in better understanding of land and sea area. The proposed architecture welcomes researchers and organizations for any type of remote sensory Big Data analysis by developing algorithms for each level of the architecture depending on their analysis requirement.

X. FUTURE ENHANCEMENT

In future work, thereare planning to extend the proposed architecture to make it compatible for Big Data analysis for all applications, e.g., sensors and social networking. We plan to divide the data into minutiae and tiny part of the data. After that, we can together them based on clients requirements. We are also planning to use the proposed architecture to perform complex analysis on earth observatory data for decision making at real time, such as earthquake prediction, Tsunami prediction, fire detection, etc.In cloud environment, the privacy preservation for data analysis, share and mining is a challenging research issue due to increasingly larger volumes of data sets, thereby requiring intensive investigation. We will investigate the adoption of our approach to the bottom-up generalization algorithms for data security. Based on the contributions herein, we plan to further explore the next step on scalable privacy preservation aware analysis and scheduling on large-scale data sets. Optimized balanced scheduling strategies are expected to be developed towards overall scalable privacy preservation aware data set scheduling.

REFERENCES

1. D. Agrawal, S. Das, and A. E. Abbadi, "Big Data and cloud computing: Current state and future opportunities," in Proc. Int. Conf. Extending Database Technol. (EDBT), 2011, pp. 530– 533.
2. J. Cohen, B. Dolan, M. Dunlap, J. M. Heller stein, and C. Welton, "Madskills: New analysis practicefor Big Data," PVLDB, vol. 2, no. 2,pp. 1481–1492, 2009.
3. J. Dean and S. Ghemawat, "Map reduce: Simplified data processing on large clusters," Common. ACM, vol. 51, no. 1, pp. 107–113, 2008.
4. H. Herodotou et al., "Starfish: A self-tuning system for Big Data analytics," inProc. 5th Int. Conf. Innovative Data Syst. Res. (CIDR), 2011, pp. 261–272.
5. K. Michael and K. W. Miller, "Big Data: New opportunities and new challenges [guest editors' introduction]," IEEE Comput., vol. 46, no.
6. pp. 22– 24, Jun. 2013. [6] C. Eaton, D. Deroos, T. Deutsch, G. Lapis, and P. C. Zikopoulos, Understanding Big Data: Analytics for Enterprise Class Hadoop and

Streaming Data. New York, NY, USA: Mc Graw-Hill, 2012.

7. R. D. Schneider, Hadoop for Dummies Special Edition. Hoboken, NJ, USA: Wiley, 2012.
8. A. Cuzzocrea, D. Saccà, and J. D. Ullman, "Big Data: A research agenda," in Proc. Int. Database Eng. Appl. Symp. (IDEAS'13), Barcelona, Spain, Oct. 09–11, 2013.
9. R. A. Schowengerdt, Remote Sensing: Models and Methods for Image Processing, 2nd ed. New York, NY, USA: Academic Press, 1997.
10. D. A. Landgrebe, Signal Theory Methods in Multispectral Remote Sensing. Hoboken, NJ, USA: Wiley, 2003.
11. C.-I. Chang, Hyperspectral Imaging: Techniques for Spectral Detection and Classification. Norwell, MA, USA: Kluwer, 2003.
12. J. A. Richards and X. Jia, Remote Sensing Digital Image Analysis: An Introduction. New York, NY, USA: Springer, 2006.
13. J. Shi, J. Wu, A. Paul, L. Jiao, and M. Gong, "Change detection in synthetic aperture radar image based on fuzzy active contour models and genetic algorithms," Math. Prob. Eng., vol. 2014, 15 pp., Apr.2014.
14. A. Paul, J. Wu, J.-F. Yang, and J. Jeong, "Gradient-based edge detection for motion estimation in H.264/AVC," IET Image Process., vol. 5, no. 4, pp. 323–327, Jun. 2011.
15. A. Paul, K. Bharanitharan, and J.-F.Wang, "Region similarity based edge detection for motion estimation in H.264/AVC," IEICE Electron. Express, vol. 7, no. 2, pp. 47–52, Jan. 2010.